# Assignment-based Subjective Questions

1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans.

- Some categorical columns like season, mnth, weekday and weatherist have numeric categorical variables with N levels. So we converted numbers into labels for visualization and later created dummy variables.
- From visualization we noticed that these columns don't have much outliers.

2) Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans.

- If you don't drop the first column of categorical n level type, then dummy variables will be correlated. That may affect the model.
- For example if we have 3 values in a column, when we create dummy variable 3 different columns are created. However, 1st column can be predicted based on other 2 columns, thus, we drop that.

3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

- Columns temp and atemp are having high correlation.

4) How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Low coef and p-values in model p-value is 0.00
- High R-squared, in model it is 80%.
- Low Prob(F-statistic) that value is low in the model

All together all these factors contribute in building good model.

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- As per the model temp, year and season_spring are driving factors of the demand for bike sharing.

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

- Linear Regression is a machine learning algorithm used to learn, derive or predict something. Regression model has a target prediction value(y) based on independent variables(x). There are two different linear regressions: they are

1) Simple Linear regression

2) Multiple Linear regression

- Equation of linear regression is close to straight line:

i) y=B0+B1x(B0 is intercept,B1 is coefficient

2. Explain the Anscombe’s quartet in detail. (3 marks)

- Anscombe’s Quartet consists of four datasets, each containing eleven (x,y) pairs. Datasets looks completely different from graph. Scattered plots are plotted across x/y plain to show coordination between variables.

3. What is Pearson’s R? (3 marks)

- Pearson’s Correlation Coefficient(Pearson’s R) helps you find out the relationship between two quantities. It gives you the association between two variables. The value of Pearson’s Correlation Coefficient can be between -1 to +1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- Scaling is a technique to standardize the independent features present in the data within a fixed range.
- It is performed during the data pre-processing to handle highly varying values.
- Standardized scaling will affect the dummy values based on original data but MinMax scaling is dummy values that will only scale between 0 and 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

  - VIF=1/1-R2
  - When R2 reaches 1 VIF become infinity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

  - *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*
  - Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.